Advances in Science and Technology
Research Journal
Volume 8, No. 24, Dec. 2014, pages 24–27
DOI: 10.12913/22998624/562

Research Article

POLISH EMOTIONAL SPEECH RECOGNITION USING ARTIFICAL NEURAL NETWORK

Paweł Powroźnik[1]

[1] Institute of the Foundations of Electrotechniqu and Electrotechnology, Faculty of Electric and Information Technologies, Lublin University of Technology, Nadbystrzycka 38A, 20-618 Lublin, Poland, e-mail: ppowroznik@gmail.com

Received: 2014.09.19
Accepted: 2014.10.25
Published: 2014.12.01

ABSTRACT
The article presents the issue of emotion recognition based on polish emotional speech analysis. The Polish database of emotional speech, prepared and shared by the Medical Electronics Division of the Lodz University of Technology, has been used for research. The following parameters extracted from sampled and normalised speech signal has been used for the analysis: energy of signal, speaker's sex, average value of speech signal and both the minimum and maximum sample value for a given signal. As an emotional state a classifier fof our layers of artificial neural network has been used. The achieved results reach 50% of accuracy. Conducted researches focused on six emotional states: a neutral state, sadness, joy, anger, fear and boredom.

Keywords: emotional speech, artificial neural network, communication.

INTRODUCTION

The recognition of the emotional state of a speaker, based on the analysis of speech signals, is a relatively new issue, however, its significance is increasing rapidly. One of the reasons of such a direction of changes is the dynamic development of systems based on the human – computer type of communication. Other applications involve processing of the speech signal, wherein speaker's emotions may play a certain significance. Among the potential applications of algorithmically modified emotional speech those associated with marketing and telephone contact with the client should be replaced [1]. Another group of applications involves the use of driver's emotional state by onboard computers. This kind of systems may be installed in vehicles, which can initiate appropriate safety procedures based on collected data [5].

The research carried out so far has been mostly based on databases in which every speech sample is matched with a specific emotional tone of voice [3]. However, the achieved results are mostly acceptable. This is due to the fact that for an average person it is possible to recognize another person's emotional state only in 60% of all cases [4].

There are several research centres in Poland which investigate the matters of emotional speech recognition (in the Polish language) [3, 5, 6]. The basic classifier applied to this type of research is the support vector machine (SVM), and the k-Nearest Neighbours algorithm (or k-NN for short).

The subject of this research is to determinate the optimal parameters for an artificial neural network allowing an effective recognition of the emotional states of a speaker. The second objective is the determination of the characteristics of the Polish emotional speech.

The article is divided into three parts. The first one characterizes the discussed matter. The second part treats the database used in the research. The third part includes the analysis of the available algorithms, research methods, and parameters for the Polish emotional speech and presents the obtained results and suggestions for improving the adapted research methods.

ANALYSIS OF ISSUES

The analysis of speech signals is connected with possessing a proper database of sound files. One of the available collections of emotional speech sound samples is the Berlin Database of Emotional Speech [7]. This database contains recordings of speech in seven emotional states, that is, a neutral state, joy, sadness, fear, anger, boredom, and disgust, prepared by 10 actors of both sexes [8]. Another group of sound files collections are databases prepared on the basis of recordings from TV and radio programs. The Lodz University of Technology has prepared and shared their own database of the Polish emotional speech. The basic problem of automatic emotional speech recognition is the choice of a proper feature vector. Descriptors, which are commonly used in this particular case do not vary from the ones used the analysis and processing of speech signals. A set of that kind of descriptors contains parameters like the signal's energy and a basic frequency [9]. Nowadays, the standards in voice recognition are *Linear Predictive Coding (*LCP), *Perceptual Linear Predictive (*PLC) [10], and *Mel-frequency Cepstrum (*MFCC) [11, 12], which are also used in the analysis of emotional speech recognition.

Based on the determined parameters of the speech signal the sets of attributes are created. These sets are then used as the starting vector for classification algorithms. As a classifier Polish emotional speech k – nearest neighbours (k – NN) algorithm [6] and support vector machine (SVM) [13] are used. However, global trends help draw the assumption of equally good performance of artificial neural networks in the analysis of the above issues [14].

The Polish database of emotional speech, prepared and shared by the Medical Electronics Division of the Lodz University of Technology, has been used for research. The collection consists of 240 recordings prepared by 8 actors: 4 women and 4 men. Each of the speakers pronounce five different sentences in 6 emotional tones, that is: boredom, fear, anger, sadness, joy, or without any emotional tone [3]. The database contains sound files in the 'wav' format sampled with 44.1 kHz frequency and the bit rate of 16 bps. The database includes the following statements: 'They have bought a new car today', 'His girlfriend is coming here by plane', 'Johnny was today at the hairdresser's', 'This lamp is on the desk today' and 'I stop to shave from today on'.

SPEECH SIGNALS PARAMETERS AND PROPOSED CLASSIFIER

The process of emotion identification based on a speech signal requires a distinction of characteristic parameters in the voice. Research carried out until now has not resulted in the determination of a uniform and universal set of features so far. As a consequence, a heuristic approach has been adopted [13]. It involves identifying, according to the signal, as many parameters describing the signal as it is possible, and selecting, by means of experimenting or using algorithms, those parameters which describe the researched matter best. Among the parameters extracted from the signal the most useful ones are: the laryngeal tone [13], energy of the signal [13], formant values, and MFCC, LPC, PLP factors [3].

ENERGY AND AVARAGE VALUE OF SPEECH SIGNAL

The research which has been carried out focuses on the energy, and the average value of the speech signal. The energy of the signal is defined as the integral of the square of the signal, that is, energy emitted with unitary resistance. For digital signals it is described by the following formula [15]:

$$Ex = \sum_{n=0}^{N} x^2(n), \tag{1}$$

where: n – sample number,
$x^2(n)$ – square value of sample signal.

The average value of the whole signal is defined in the following way [15]:

$$\overline{x_N} = \lim_{n\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n),$$

where: $x(n)$ – value of n-sample,
N – total mumber of samples.

For the sake of the research, before the abovementioned parameters have been set, the values of particular samples have undergone the process of normalization. A standard algorithm of signal processing is based on three basic stages: preparation of the data set, the designation of the feature vector and classification. Thanks to the high quality of recordings the first stage has been limited to normalization process.

The second stage is the determination of the feature vector describing the analyzed matter as precisely as possible. In this study, this stage was

limited to five elements: the energy of the signal, the average value of the whole signal, the sex of the speaker (1 – woman, 0 – man), and both the minimum and maximum sample value for a given signal. This set is the entry vector in a neural network. The last stage is classification. A five entry neuron network has been proposed as a classifier, twelve neurons in the first, hidden layer, six in the second one, and six entry neurons. The model of the network is depicted in Figure 1.

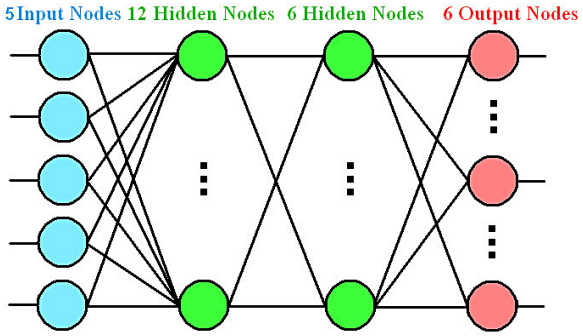

Fig. 1. The model of used Artificial Neural Network

The sigmoidal function has been used as an activation function. The research has been carried out in MatLab, where the network has been trained with the backpropagation with momentum and factor adaptation (traingdx). The learning process ended with either the achievement of the given number of epochs, that is, 1500 in the analyzed cases, or the achievement of a normalized result different from the expected one by no more than 0.1. The achieved results have been depicted in Figure 2. The confusion matrix has been shown in Table 1.

CONCLUSIONS

The recognition of emotions in speech signal is a difficult task, and the achieved results are far from ideal. Research carried out by Swiss scientists show, that the evaluation of an emotional state is difficult even for a human being. Nevertheless, it is possible to improve the achieved results [4]. The first step is to expand the fea-

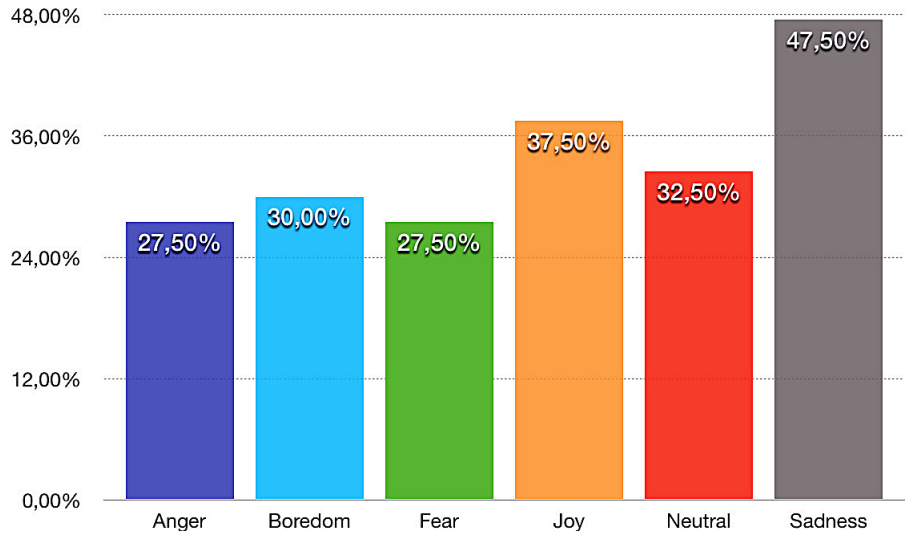

Fig. 2. The effectiveness of the recognition of individual emotional states

Table 1. Confusion matrix

		Actual class					
		Anger	**Boredom**	**Fear**	**Joy**	**Neutral**	**Sadness**
Predicted class	**Anger**	**11**	5	10	12	1	1
	Boredom	4	**12**	6	1	12	5
	Fear	10	8	**11**	6	3	2
	Joy	11	1	8	**15**	4	1
	Neutral	4	10	4	1	**13**	8
	Sadness	1	8	6	1	5	**19**

ture vector. The signal energy and its average value are not enough to achieve good results. It is absolutely necessary to expand above set of parameters with MFCC and LCP parameters. Constructed artificial neural network (ANN) has also failed to accomplish the expectations. It is necessary to expand the ANN or to transform it into, for instance, the Kohonen network. The effectiveness of emotion recognition can be also increased by combining a voice analysis system with semantic analysis [16]. A natural direction of development is, first and foremost, the application and testing of the suggested solutions in both the process of abstracting signal and classifier parameters.

REFERENCES

1. Ramakrishnan S.: Recognition of emotion from speech: A review. Speech Enhancement, Modeling and Recognition – Algorithms and Applications, March 2012.
2. Kamaruddin N., Wahab A.: Driver behavior analysis through speech emotion understanding. Intelligent Vehicles Symposium (IV), IEEE, 2010, 238–243.
3. Kamińska D., Pelikant A.: Zastosowanie multimedialnej klasyfikacji w rozpoznawaniu stanów emocjonalnych na podstawie mowy spontanicznej. IAPGOŚ 03, 2012.
4. Scherer K.: Vocal communication of emotions: A review of research paradigms in speech communication, 40, 2003, 227–256.
5. Database of Polish Emotional Speech, available: http://www.eletel.p.lodz.pl/bronakowski/med_catalog/ (Accessed 10.08.2014).
6. Ślot K., Rozpoznawanie biometryczne, WKiŁ, Warszawa, 2010.
7. Berlin Database of Emotional Speech, available: http://www.expressive-speech.net/ (Accessed 10.08.2014).
8. Polzehl T., Schmitt A., Metze F.: Approaching multi-lingual emotion recognition – from speech – on language dependency of acoustic/prosodic features for anger recognition. Proc. of Speech Prosody, Chicago 2010.
9. Yeqing Y., Tao T.: An new speech recognition method based on prosodic analysis and SVM in Zhuang language. Proc. 2011 International Conference on Mechatronic Science, Electric Engineering and Computer, 2011, 1209–1212.
10. Shauka A., Chen K.: Emotional state recognition from speech via soft-competition on different acoustic representations. Proc. Neural Networks (IJCNN), 2011, 1910–1917.
11. Plutchik R.: The nature of emotion. American Scientist, Vol. 89, July-August 2001, 344–350.
12. Niewiadomy D., Pelikant A.: Implementation of isolated words boundaries recognition. Proc. XII International Conference System Modeling and Control SMC, Zakopane 2006.
13. Janicki A., Turkot M.: Rozpoznawanie stanu emocjonalnego mówcy z wykorzystaniem maszyny wektorów wspierających (SVM). KSTiT, Bydgoszcz 2008.
14. Soltani K., Ainon R.: Speech emotion detection based on neural networks. Proc. Signal Processing and Its Applications, 2007, 1–3.
15. Zieliński T.: Cyfrowe przetwarzanie sygnałów. Od teorii do zastosowań. WKŁ 2009.
16. Wang Y., Guan L.: Recognizing human emotional state from audiovisual signals. Proc. IEEE Transactions on multimedia, Vol. 10, 2008, 659–668.